Single-cell transcriptional profiling: a window into embryonic cell-type specification

Blanca Pijuan-Sala[1], Carolina Guibentif[1] & Berthold Göttgens[1]

[1]Wellcome and MRC Cambridge Stem Cell Institute, and the Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK.

Correspondence to B.G. bg200@cam.ac.uk

Abstract

During mammalian embryonic development, a single fertilized egg cell will proliferate and differentiate into all the cell lineages and cell types that eventually form the adult organism. Cell lineage diversification involves repeated cell fate choices that ultimately occur at the level of the individual cell rather than at the cell-population level. Recent improvements in single-cell technologies are beginning to transform our understanding of mammalian development, not only by overcoming the limitations presented by the extremely low cell numbers of early embryos, but also by enabling the study of cell-fate specification in greater detail. In this Review, we first discuss recent advances in single-cell transcriptomics and imaging, and provide a brief outline of current bioinformatics methods available to analyze the resulting data. We then discuss how these techniques have contributed to our understanding of pre-implantation and early post-implantation development, and of *in vitro* pluripotency. Finally, we overview the current challenges facing single-cell research and highlight the latest advances and potential future avenues.

Introduction

During the early stages of mammalian embryonic development, the fertilized egg will give rise to all the cell types that form the adult organism through sequential rounds of proliferation and differentiation. The robustness of this process indicates that development is governed by a set of principles that guide each cell to commit to becoming a particular cell type. But what are these principles? How do cells consistently activate certain molecular programs and not others? There has been extensive research trying to decipher these principles by assessing how genes regulate each other. Conventional experimental approaches, however, required large homogeneous cell populations and were therefore not applicable to the early mammalian embryo, which is composed of very few cells that become increasingly heterogeneous as the embryo develops**.**

With the advent of high-throughput single-cell sequencing and single-molecule imaging technologies, it is now possible to generate single-cell molecular profiles at a large scale, and to quantify single-cell gene expression more accurately in defined multicellular tissues. Since cell fate decisions are ultimately made by individual cells responding to external stimuli, these new transcriptional technologies are now making a great impact on our understanding of early embryonic development. In this Review, we summarize the recent advances in single-cell transcriptional technologies with a particular focus on single-cell transcriptomics and imaging, and provide a brief outline of the current bioinformatics methods to analyze the resulting data. We then focus on how these techniques have contributed to our understanding of *in vivo* pre-implantation and early post-implantation mouse development, and of *in vitro* pluripotency. Finally, we overview the current challenges of using these technologies and highlight some future avenues for the field.

Single-cell technologies

One of the most important discoveries in the single-cell field was the finding that single cells from organs or cell types that were previously considered to be homogeneous are often functionally and molecularly

heterogeneous. This initially became evident using flow cytometry, which is one of the first high-throughput single-cell analysis approaches that now permits the assessment of up to 40 cell-surface marker expression profiles in individual cells when combined with elemental mass spectrometry (reviewed in [1]). The establishment of fluorescence-activated cell sorting has also enabled sorting individual cells, or molecularly defined cell populations, into single wells to evaluate them separately[2]. As outlined below, more recent technologies now permit the simultaneous measurement of many parameters of individual cells, which has contributed to the discovery of novel regulators of cell differentiation and developmental processes.

Single-cell transcriptomics.

Single-cell transcriptional analyses were pioneered with the amplification of polyadenylated RNA from individual hematopoietic cells to measure relative abundances between transcripts[3]. However, this was a qualitative approach and it was not until the development of quantitative real-time PCR (qPCR)[4,5] that single-cell qPCR (sc-qPCR) became feasible. Since then, significant progress in automation and the development of microfluidics systems has allowed measuring the expression of several genes in several single cells in one run[6]. This targeted analysis of known genes has been instrumental in characterizing molecular programs of multilineage cell progenitors, which seem to co-express genes characteristic of the different mature cell types they will give rise to[7,8]. It also essential for the study specific signaling pathways or transcription factor networks in developmental cell populations, ranging from the zygote to the differentiation of particular cell lineages, such as blood or endothelial cells[6,9,10]. However, to identify novel factors and signaling pathways at play during embryonic development, genome-wide, transcriptomics approaches are required.

Transcriptomics technologies began with the development of single-cell microarrays to study gene expression in the mouse blastocyst [11]. Using this strategy, embryonic single cells were molecularly classified into two groups, epiblast and primitive endoderm , which are two lineages known to segregate at the blastocyst stage[11] (BOX 1). Although microarrays could already distinguish between cells of different lineages, they are restricted to analyzing characterized transcripts, from which the probes used to measure transcript abundances are derived.

The need for a more comprehensive tool led to the development of the first single-cell RNA-seq (scRNA-seq) protocol in 2009[12]. This method was able to detect a considerably higher number of genes in blastocysts compared to single-cell microarrays and, therefore, a more thorough transcriptional characterization of this developmental stage could be achieved[12]. Since then, several optimized protocols have been developed, including Smart-Seq2 and cell expression by linear amplification and sequencing 2 (CEL-seq2)[13–16] (BOX 2). These improved methods achieve a higher transcript coverage compared to early scRNA-seq protocols and are considered the gold standards in the field. However, they still require the isolation of individual cells into single wells, and are thus designated ‘well-based’ methods. Such methods utilize strategies such as fluorescence-activated cell sorting or microdissection, which restrict cell throughput.

To transcriptionally characterize whole organs or embryonic stages at the single-cell level in a genome-wide manner, ‘droplet-based’ microfluidics protocols are a suitable alternative[17,18] (BOX 2). These allow processing thousands of cells in one experiment, thus contributing to the discovery of rare subpopulations within the sample or to the characterization of its molecular diversity. This methodology has recently been used to transcriptionally characterize several biological systems, such as adult retinal cell populations[18], mouse embryonic stem cells (mESCs)[17] and the entire mouse embryo at day 8.25 of development (E8.25)[19]. However, droplet-based techniques also have some limitations, such as the higher risk of generating data from cell doublets, a transcript coverage bias towards 3’ ends and being less sensitive than well-based methods due to capturing fewer unique transcripts. Therefore, well-based methods remain a good choice for answering specific questions that require a better transcriptome coverage and/or a relatively low cell number.

Whether using a well-based or a droplet-based method, single-cell transcriptomics protocols require a pre-amplification step to obtain enough material for the downstream detection method, and, therefore, the raw data obtained in these methods reflect the relative abundances of the amplified cDNAs instead of the original amounts of RNAs in the sample. Since PCR amplification occurs in a non-linear fashion, quantitative biases could be easily introduced in the resulting gene counts. Therefore, optimizing the number of PCR cycles in order to keep every reaction within the exponential phase by performing a preliminary set of tests, can help reducing the risk of bias. Alternatively, a more precise method to quantify RNA reliably is the addition of molecular labels called unique molecular identifiers before amplifying the biological material[20]. By tagging each transcript individually, the absolute numbers of RNA molecules in each cell can be estimated,

thus minimizing amplification noise.

Due to the small amount of starting material, scRNA-seq data is prone to contain technical noise, which negatively correlates with transcript abundance[21]. Moreover, batch effects can confound both data analysis and display. These could be caused by several experimental conditions, ranging from using different reagent batches to performing the experiment on different days. To counteract these problems, normalization strategies and ways to bioinformatically infer the highly variable genes have been developed[21–25]. Once the raw sequencing counts are processed, current single-cell transcriptomics studies use dimensionality-reduction approaches such as principal component analysis, t-distributed stochastic neighbour embedding (t-SNE)[26] and diffusion maps[27,28] to visualize the data (FIG. 1A). In contrast to principal component analysis, t-SNE and diffusion maps are able to capture the non-linear nature of biological processes; t-SNE is particularly useful for identifying the main transcriptionally-defined cell groups that exist within the dataset. For instance, this strategy has been applied to resolve the biological structure of a scRNA-seq dataset composed of cells collected during gastrulation (BOX 1). This revealed the existence of ten distinct cell groups, which included populations of epiblast cells, mesoderm cells, blood and endothelial progenitor cells and differentiated cells[29]. Importantly, the distances between the different groups visualized in t-SNE plots should not be considered a biological representation of relatedness. Coupled with cell clustering strategies, which help reveal the prevalent gene expression profiles within the dataset, t-SNE has allowed the transcriptional characterization of embryonic cell populations from single cells ranging between the early pre-implantation stages, gastrulation and organogenesis, as well as the transcriptional characterization of developing tissues, such as the heart and the blood[10,19,29–32].

Diffusion maps aim to display the data on a continuous manifold, where nearby cells have higher transcriptional similarities, and therefore are commonly used to visualize biological processes such as cell differentiation. Diffusion maps are often coupled with pseudotime algorithms, such as Diffusion Pseudotime[9], Wishbone[33], Monocle[34] and Monocle2[25] (FIG. 1B). These analyses are used to infer a sequential progression by computationally ordering the cells along a coordinate that reflects the developmental path and, therefore, allow the study of gene expression dynamics along this trajectory. This has been applied to visualize the differentiation of many cell lineages, including those of blood cells[10,28], lymphoid cells[33,35] and mESCs[9], and helped to discover novel regulators of these processes. Importantly, pseudotime analyses assume that these biological processes are continuous, and thus position cells that share a similar transcriptional profile close in the developmental path. Consequently, they may not be accurate if the biological system behaves in an oscillatory manner or undergoes fast state transitions, thereby necessitating experimental validation. In such cases, previous knowledge of, for instance, specific genes known to have gradient expression along the trajectory of the studied process, can help infer the developmental progression from snapshot data. This approach has recently been applied to predict genes with wave-like expression during mouse somitogenesis from cells at E8.25, and was subsequently validated using microdissection[19].

High-throughput single-molecule imaging

A major limitation of single-cell transcriptomics is the lack of spatial and temporal resolution. Many studies have therefore used imaging technologies to understand how gene expression occurs and is orchestrated in single cells across a given tissue. Although imaging has generally been restricted by the number of transcripts one can assess due to the limited range of discernible dyes, it is now possible to spatially resolve the expression of about 10^3 genes simultaneously in single cells using single-molecule RNA fluorescent in situ hybridization (sm-FISH)[36–38]. This is feasible due to a combination of different sequence-specific probes that target each transcript multiple times, thus generating a unique transcript barcode, which can be resolved with microscopy (FIG. 1C). sm-FISH has been applied in mESCs to describe their molecular heterogeneity[39]. This technique has also been used together with a CRISPR–Cas9 tool to establish a system called MEMOIR that records lineage information in single mESCs[40]. Importantly, sm-FISH is a targeted approach, which requires costly pre-designed probes to study all the genes of interest. Moreover, the technique currently requires bespoke imaging equipment and software, which, at least in the near future, will limit its use. Recently, a more indiscriminate strategy called fluorescent in situ sequencing (FISSEQ) has been established[41], which enables unsupervised sequencing of single-cell transcripts *in situ*. However, its current detection limit of about 200 transcripts per cell and its high cost are considerable drawbacks.

The aforementioned methods provide temporal snapshots. Therefore, to validate the dynamics of a particular process, it is important to resort to live imaging techniques, which can track gene expression changes over time. In the past few years, major improvements enabled the detection of single molecules in living cells. For example, RNA can be labeled using the MS2 and PP7 systems, which consist of RNA hairpin sequences

fused in tandem to the transcript of interest, and which are bound by a fluorescent protein with high specificity[42,43]. However, in practice, most live imaging studies to date have used fluorescent reporters to track the changes in protein levels instead of transcript levels. This has produced controversial results regarding cell fate specification. For instance, gene regulation inferences from single-cell transcriptional studies in mice have suggested that during hematopoietic differentiation, the transcription factors Gata1 and Pu.1 are the master regulators of cell-fate choice between the erythroid and myeloid lineages, respectively, by cross-repressing each other[44]. Recently, this model has been challenged by data obtained from tracking Gata1 and Pu.1 proteins using live imaging methods, which showed that Gata1 protein levels are independent of lineage specification[45]. This is a paradigmatic example of the importance of using complementary-analysis approaches to validate the models inferred by single-cell transcriptomics studies.

Heterogeneity in cell fate decisions

One of the most fascinating and still unanswered questions in developmental research is how single cells consistently commit to particular lineages. Are the lineages molecularly primed at an early stage, where the cells are functionally indistinguishable by current methods, or do they become restricted just prior to their functional segregation? Single-cell transcriptional profiling addressed this question in the context of the first three lineage-branching points of embryonic development: (1) from a totipotent cell to the inner cell mass (ICM) and trophectoderm; (2) from the ICM to the epiblast and the primitive endoderm; and (3) from the epiblast to the ectoderm, mesoderm and endoderm. In this section, we will discuss recent findings achieved using single-cell experiments to understand these cell-fate decisions.

From totipotency to the inner cell mass and trophectoderm

The first lineage segregation in the mouse embryo will lead to the generation of the so-called ICM and the trophectoderm (BOX 1). This distinction becomes morphologically and molecularly apparent at the blastocyst stage, where cells are divided into round-shaped ICM cells expressing epiblast and primitive endoderm markers such as the transcription factors Nanog, Oct4 (also known as Pou5f1) and Gata6, and to relatively flat outer trophectoderm cells with high levels of the trophectoderm marker caudal-type homeobox protein 2 (Cdx2)[46–51]. Although these lineages have been widely characterized, how this segregation occurs remains controversial.

In the past few years, some studies have reported transcriptional heterogeneity between very early blastomeres of many factors that eventually confer the ICM or trophectoderm identities[52,53]. How this early heterogeneity affects cell fate decisions is currently under debate with two contrasting hypotheses. The equivalence hypothesis argues that individual cells are not pre-patterned until the 8-cell stage, when the embryo undergoes a series of symmetric and asymmetric divisions that would drive the internalization of the cells that will become part of the ICM[52,54–56,57–60]. Under this hypothesis, the heterogeneity seen in early stages is explained as random fluctuations of gene expression. By contrast, the asymmetric hypothesis proposes that partitioning inequities occur during cell divisions, which lead to cellular components being unevenly distributed in the daughter cells[61–65]. This asymmetry becomes amplified through consecutive divisions and eventually leads to lineage segregation. This hypothesis not only implies that cells would start showing some preferences towards particular lineages before the 8-cell stage, but that this asymmetry would be fairly reproducible between cells when comparing different embryos.

Imaging experiments assessing the behavior of some ICM and trophectoderm markers, such as Nanog, Oct4 and Cdx2, have shown that these are expressed in a random and uncorrelated manner before lineage segregation is apparent at the 16-cell stage[52]. Although this would support the equivalence hypothesis, these factors could alternatively be essential only once the lineages are established, and some other genes could be responsible for an early cell pre-patterning. To address whether other molecular candidates exist before the 16-cell stage, genome-wide transcriptome analyses are key. Recently, several groups have applied scRNA-seq on cells collected during these stages and have shown that some genes are already expressed in a bimodal manner early on, and their transcription is positively or negatively inter-correlated, which is unlikely to be compatible with random fluctuations[66,67].

Gene expression heterogeneity between cells seems to start at the 2-cell stage and increases over time, with significant differences at the 2-to-4 and 4-to-8 cell transitions[53,66,67] (FIG. 2A). Particularly, the transitions where the highest changes in gene expression were observed correspond to the stages where the maternal-to-zygotic transition occurs[6866]. Since this process consists of the degradation of maternal

transcripts and transcription activation of the zygotic genome, it is expected to generate a bimodal distribution of the levels of the transcripts involved. Although the timing of events and some expression profiles differ from mice[69–71], similar findings were made in the human setting, where single-cell transcriptomic studies have also shown that the greatest gene expression changes take place during the maternal-to-zygotic transition[70–72], supporting the hypothesis that molecular segregation occurs shortly after fertilization. However, these results do not prove that this transcriptional segregation directly determines cell fates. Of note, recent evidence suggests that the molecular variability seen during earlier stages may bias cells towards undergoing either asymmetric or symmetric divisions during the 8-to-16 and 16-to-32 cell stages, which, in turn, would influence cell fate. The transcription factor SRY-box 21 (*Sox21*) was reported as heterogeneously expressed at the 4-cell stage. When knocked down in one of the blastocysts at the 2-cell stage, reduction of *Sox21* reduced the cell's probability to undergo asymmetric divisions, and enhanced the contribution towards the trophectoderm[53] (FIG. 2A). These results support a model where early transcriptional heterogeneity of some genes influences the type of cell division and fate from very early stages. However, the evidence provided for these conclusions was indirect, with no direct demonstration that a consistent inter-blastomere segregation of Sox21 mRNA and protein content continues to exist after the 4-cell stage, and that this in turn correlates with lineage segregation at the 8-to-16 cell stage.

Regarding the question of how asymmetric divisions would then influence cell fate, recent studies have related the fate of daughter cells to different timings of cell division, as well as to their exposure to different physical conditions, which then regulate the internalization of the cells that will subsequently become ICM[58,58,60,73]. Such physical conditions include differences in membrane tension, inheritance of the apical pole, and contractility of the ICM (reviewed in [74,75]).

Specification of the inner cell mass to epiblast or primitive endoderm

This first decision, of specification into ICM or trophectoderm, is partly coupled with the second fate choice, where the inner cells will become either epiblast or primitive endoderm. It was first thought that cells become segregated to either lineage depending on their localization within the ICM. However, in 2006, it was demonstrated that cells first become molecularly specified in a 'salt-and-pepper' manner within the ICM and later organize themselves to achieve their final configuration: inner epiblast cells surrounded by primitive endoderm[76]. But how is this achieved? How do ICM cells choose between an epiblast and a primitive endoderm identity?

Once at the morula stage (8 cells in the mouse), cells that will form the ICM become internalized during the following two rounds of asymmetric divisions: from 8-to-16 cells and from 16-to-32 cells [57–60]. Some evidence suggests that the round of division where the ICM cells become internalized influences their epiblast *vs.* primitive endoderm fate[77,78] (FIG. 2B). More specifically, cells that internalize during the first round of asymmetric divisions tend to become epiblast whereas a later internalization leads to a higher probability of becoming primitive endoderm. However, whether this is the case has been debated[79] and an alternative mechanism, consisting of a random specification towards epiblast or primitive endoderm has been proposed (FIG. 2B). Whichever mechanisms may be at play (see below), single-cell transcriptional studies using sc-qPCR and single-cell microarrays have shown that during early stages, ICM cells are highly heterogeneous and co-express different epiblast and primitive endoderm markers, which eventually adopt a mutually exclusive pattern[6,80,81].

The functions of the transcription factors Nanog and Gata6 and of fibroblast growth factor (FGF) signaling are key to ICM segregation[6,76,78–80,82–84] (FIG. 2C). Single-cell imaging analyses have shown that Nanog and Gata6 become heterogeneously expressed at the 8-to-16 cell stage, with some cells expressing both. Slightly later, these factors are expressed in a mutually exclusive manner, which marks the initiation of lineage segregation[76,82]. To induce epiblast formation, Nanog activity is required to suppress Gata6 expression[82]. However, the sole expression of Gata6 in Nanog-negative cells is not sufficient to impose the primitive endoderm identity and FGF signaling is also required[82]. At the 16-to-32 cell stage (E3.25), *Fgf4* becomes expressed and secreted by a cell subpopulation that already expresses Nanog, and is key for the paracrine induction of the primitive endoderm factors Gata4 and Sox17 in Gata6-expressing cells, thus committing the cells to a primitive endoderm identity[82]. Notably, rarely a few primitive endoderm-biased cells appear to turn into epiblast cells by re-activating *Nanog* expression during the initial phases of lineage segregation, indicating that fate change is still possible. However, once the epiblast and primitive endoderm fates have been established following FGF signaling, fate switch is absent[81]. Infrequent cell fate switching has also been reported *in vitro* in ESCs.

Single-cell transcriptome analyses have shown a strong negative correlation between *Fgf4* and *Fgfr2* expression, suggesting that the direct interaction between Fgf4 and Fgfr2 expressed in different cells could be important for the ICM cell fate decision[6,78,80,82]. However, whereas mutations in Fgf4 completely disrupt primitive endoderm formation, mutations in Fgfr2 are not sufficient to abolish this lineage[83,84]. During these early stages of development, *Fgfr1* is also broadly expressed in the ICM, suggesting it has a role in ICM segregation[80]. This hypothesis, which originated from single-cell transcriptional studies, was recently confirmed by two single-cell imaging studies that showed Fgfr1 is required for ICM specification[83,84]. Thus, the repression of Gata6 by Nanog initiates ICM lineage segregation, and this is subsequently coupled with Fgf4 activity, which completes the epiblast and primitive endoderm specification through its direct interactions with the receptors Fgfr1 and Fgfr2 (FIG. 2C)**.**

Post-implantation epiblast cells: committing to mesendodermal fates

Once the epiblast and primitive endoderm lineages have segregated, and after the embryo implants into the uterus and undergoes a series of morphological changes (see [85,86] for a review), gastrulation begins (BOX 1). During this process, the three germ layers – ectoderm, mesoderm and endoderm – are formed. While mesodermal and endodermal cells arise from epiblast cells egressing through the primitive streak, the remaining epiblast cells will become the ectoderm[87]. A series of transplantation assays performed in the 1990s showed that epiblast cells do not appear to be committed to a specific lineage during this stage; rather, they seem to remain pluripotent and to adopt a particular fate depending on when and where they egress through the primitive streak[88]. However, fate-mapping studies have shown that cells residing in particular regions within the epiblast have a preference to contribute to a specific tissue[87]. This would suggest that cells are already biased in the epiblast but, as morphogenetic movements are highly organized at this stage, the starting position rather than any specific transcriptome differences could be the determining factor of epiblast contribution to later-forming tissues.

Post-implantation epiblast cells were recently profiled using single-cell transcriptomics to determine whether they exhibit any priming towards a particular lineage[29,89]. These studies have shown that post-implantation epiblast cells are molecularly heterogeneous, but they do not appear to express markers indicative of early segregation to mature lineages. Interestingly however, both studies observed some cells that co-expressed transcription factors known as mesoderm and endoderm markers, such as the *T* gene (encoding the Brachyury protein) and hepatocyte nuclear factor 3β (also known as *Foxa2*), indicative of a mesendodermal progenitor identity[90,91]. This subpopulation seems to appear as early as E5.5[89], when the primitive streak is not morphologically visible, suggesting that this could be one of the earliest populations committing towards a mesendodermal fate. Moreover, this also suggests that molecular changes precede morphological changes during gastrulation, although spatial information would be needed to confirm this (FIG. 2D).

Since the E6.5 post-implantation epiblast population appeared to contain the transition from pluripotent epiblast to mesoderm and definitive endoderm, cells were examined for genes whose expression positively or negatively correlated with the expression levels of *T*[29]. Although some of the resulting differentially expressed genes have been already reported in the literature, such as the transcription factor *Mixl1*, these analyses also revealed some potentially novel factors, like solute carrier family 35 member D3, which may have an important role in the generation of mesoderm from epiblast cells. Their further characterization could therefore provide insights into the regulation of epiblast commitment towards mesendoderm.

Gastrulation

During gastrulation, as the cells egress through the primitive streak, both the timing and anterior–posterior location along the primitive streak have been shown to influence their lineage fate[85,88]. The molecular mechanisms establishing this diverse range of progenitors along the streak and the transcription factors regulating mesendodermal cell differentiation towards mature cell types are still obscure, but single-cell transcriptional analyses have contributed to their characterization.

Heterogeneity within the mesodermal population

To investigate the maturation of a specific lineage from the mesoderm, it is crucial to have a good

understanding of the molecular properties of that particular mesodermal progenitor so that we can identify it. Nevertheless, assigning a unique combination of markers to define specific progenitor populations has been difficult and most currently used markers are expressed in several lineages. Functional analyses of progenitor populations defined by ambiguous markers contributed to the attribution of multipotency to mesodermal cells, but a single-cell assessment has revealed that this may not always be the case. For instance, mesodermal cells expressing the kinase vascular endothelial growth factor receptor 2 (also known as Flk1) can give rise to blood, endothelial and smooth muscle cells[92–94]; however, these cells are highly heterogeneous at the molecular level[29] (FIG. 3A). Another case is the transcription factor ETS translocation variant 2 (Etv2), which is required for endothelial, endocardial and blood development[95,96]. Using scRNA- seq, *Etv2*-expressing cells were shown to be broadly transcriptionally divided into mesodermal, blood, endocardial and endothelial subpopulations, suggesting that, in addition to immature multipotent mesodermal progenitors, unipotent progenitors may also exist[97].

The reliance on markers could also be misleading. For instance, during heart development, cardiac cells form two different structures: the first heart field and the second heart field. Both populations appear to arise from mesoderm posterior protein 1 (Mesp1)-expressing mesodermal progenitors, but whether a common progenitor or different cardiac progenitors exist that give rise to both fields could not be assessed with population assays[98]. When performing lineage tracing of single clones, each mesodermal clone contributed to either the first heart field or the second heart field, showing that two distinct Mesp1- expressing mesodermal progenitor populations exist[99,100]. Moreover, sc-qPCR revealed that these populations have different molecular characteristics[100], suggesting that transcriptional characterization of progenitor populations at the single-cell level will contribute to resolving their heterogeneous differentiation potential.

From mesodermal progenitors to differentiated cell types

Following gastrulation, cells are thought to receive cues that promote changes in gene expression to drive specification towards a particular lineage. To help discover potential regulators of these processes, computational methods were recently applied to gene expression data obtained from individual cells to study the origins of blood cells[10,19,29]. During embryonic development, mesodermal progenitors give rise to blood through multiple waves, the first two occurring in the yolk sac (FIG. 3B). To investigate the first wave of blood differentiation from mesoderm, single cells positive for Flk1 or CD41, proteins that are expressed during blood development, were profiled using scRNA-seq and, applying pseudotime inferences, a differentiation trajectory was proposed[29]. Notably, the reconstructed trajectory identified some ChIP-seq-verified target genes of the blood-differentiation regulator Gata1, such as Nuclear factor, erythroid derived 2 (*Nfe2*), as being upregulated after the onset Gata1 expression, thereby validating the trajectory[29].

To discover novel factors involved in the emergence of definitive blood within the yolk sac, a Boolean algorithm was applied on sc-qPCR data from cells ranging from uncommitted mesoderm to the appearance of definitive blood in the yolk sac, leading to a model of the underlying gene regulatory network[10]. The produced model predicted the previously unrecognized regulation of *Erg* by the Hox and Sox transcription factors, which was then experimentally validated. Moreover, it provided a platform for performing *in silico* perturbations, which predicted the requirement of *Sox7* downregulation for blood emergence in the yolk sac during development. This novel finding was also functionally validated using transgenic mouse embryos with enforced *Sox7* expression, showing a marked defect in blood maturation. Further characterization of the emergence of definitive blood using single-cell transcriptomics has recently guided the discovery of the leukotriene pathway as a previously unrecognized regulator of yolk sac definitive hematopoiesis[19]. Together, these studies demonstrated how single-cell transcriptional measurements not only provide a snapshot of the molecular profiles underlying blood emergence, but also enable the prediction of the molecular mechanisms and gene regulatory networks controlling these developmental processes.

Pluripotency in vitro: stem cells

Although *in vivo* pre-implantation studies have provided insights into how cells become segregated into the epiblast and the primitive endoderm lineages, the transient nature of pluripotent ICM cells limit their use for investigating the molecular process behind the maintenance of, the entry to and the exit from pluripotency. Understanding this process requires a cell system that is easy to manipulate, able to both self-renew and differentiate, and also that gives the flexibility to follow gene expression dynamics throughout the transition

between states.

Defining subpopulations within mouse embryonic stem cells

Mouse ESCs have been one of the most popular tools used to understand pluripotency. mESCs were obtained from the ICM in 1981 and were first grown in culture on a feeder cell layer[101]. Advancements in the understanding of their needs led to the two most popular feeder-free culturing strategies, namely culturing in fetal bovine serum and leukemia inhibitory factor (serum–LIF), or in 2i, which is a serum-free medium supplemented with the GSK3β and Mek inhibitors (see [102] for a review). DNA methylation profiling as well as single-cell transcriptional studies have shown that mESCs can be divided into at least two major subpopulations: naïve pluripotent cells expressing pluripotency factors, and epiblast-primed cells, which express higher levels of keratins and actins[24,39,103]. Whereas mESCs cultured in serum–LIF are highly heterogeneous, 2i conditions tend to restrict them to a naïve state[39,103,104].

Nanog is a pluripotency factor that has been widely used to characterize pluripotent states in culture, with high Nanog expression being characteristic of the naïve state. Transitions between the Nanog-high and Nanog-low populations have been reported though these are infrequent, which makes the study of the mechanisms behind the entry to and the exit from pluripotency difficult[39,105–107]. As seen *in vivo*, Nanog expression levels also appear to be heterogeneous within cell populations grown over time in serum–LIF media, which suggests that cell-state change from the naïve to the primed state or vice versa could occur when reaching the lower or upper levels of Nanog expression, respectively. Importantly, this heterogeneity does not seem to be completely random and a mechanism for cell-state maintenance or switch may exist (FIG. 4A). Nanog expression appears to correlate with the cell cycle[39,105,106]: following cell division, as cells progress through the cell cycle, Nanog levels gradually rise until the next cell division, where its levels abruptly drop. Transcription noise leading to a higher transcription rate and an extended cell cycle are two non-exclusive hypotheses for increasing Nanog expression to levels that would maintain pluripotency or promote cell-state switch from the primed to the pluripotent state, as both would cause Nanog to gradually accumulate in the cell during consecutive cell cycles[39] (FIG. 4A). Consistent with this hypothesis, mESCs grown in 2i conditions and mostly composed of cells in the naïve state, have higher transcription rate and a longer cell cycle[24,39], which would favor the Nanog-high pluripotent state[39,106].

Recently, single-cell imaging has also allowed the discovery of a rare cell subpopulation within cultured mESCs, presenting some molecular features resembling the embryonic 2-cell (2C) stage[108]. This 2C-like population expresses *Gag* and mouse endogenous retrovirus-L (MERVL), which are two retroviral genes that are transiently expressed *in vivo* during the 2-cell stage[108], as well as zinc finger and SCAN domain containing protein 4C (*Zscan4*), which is important for genome stability[109]. Initial single-cell transcriptome profiling clustered the 2C-like cells with cells from the blastocyst stage instead of with the 2-cell stage[24]. However, a recent single-cell transcriptome re-evaluation has demonstrated that the genes upregulated in the *MERVL*- and *Zscan4*-expressing cells are those activated during the MZT, which occurs around the 2–4 cell stage[68,110]. Consistent with being a distinct cell type and in contrast to ESCs or epiblast cells, 2C-like cells present an open chromatin profile and their DNA is hypomethylated, thereby resembling the *in vivo* 2-cell stage[110]. Further supporting their 2-cell identity, 2C-like cells appear to be totipotent, as they are able to colonize both embryonic and extra-embryonic tissues upon injection into mouse blastocysts[108].

Induced pluripotent stem cells: reprogramming towards the pluripotent state

The discovery of induced pluripotency by Shinya Yamanaka and colleagues, who reported that the mESC properties of pluripotency and self-renewal can be imparted into mature somatic cells upon over- expression of four transcription factors — Oct4, Sox2, Krüppel-like factor 4 (Klf4) and Myc (OSKM) — has reshaped our understanding of pluripotency regulation[111,112]. Induced pluripotent stem cells (iPSCs) not only represent a promising technology for regenerative medicine, but they also constitute a formidable tool to unravel how pluripotency and differentiated lineages are established. Since the seminal reports using retroviral-mediated overexpression of the OSKM factors for iPSC generation[111], considerable effort has been invested towards improving the initially inefficient rates of cell reprogramming. For example, FACS and mass spectrometry were used to characterize and purify cell-surface marker-defined populations at intermediate stages of reprogramming[113,114], but the current proportion of iPSCs produced is still low.

Single-cell qPCR of human and mouse cells has shown that during the early stages of reprogramming, there is an increase in transcriptional heterogeneity within the culture, followed by a gradual decline over

time[115,116]. This initial variation could be attributed to differences in the expression levels of different OSKM factors, as individual cells appear to overexpress the four retroviral constructs at different stoichiometries during the early stages of reporgramming[117]. When evaluating gene expression distributions, dynamically expressed genes tend to start with a bimodal profile at the population level, and to later switch to a unimodal distribution [116,118,119]. A similar scenario is also seen during the early stages of mESC differentiation[17], which may suggest that, following external cues, cells start modifying their expression profiles at different rates. This initial heterogeneity has formally been described as a stochastic process, governed by an ordered set of probabilistic events that regulate the expression of each gene independently[115,116]. Supporting this model, single-cell live imaging and mass cytometry studies have also shown that, upon OSKM expression, cells start dividing at different rates: cells that quickly proliferate and reduce their cellular area are more likely to become iPSCs[120], whereas those that with a lower proliferative profile tend to revert to the original differentiated state[117] (FIG. 4B). This indicates that the speed ofproliferation is a limiting step during the early stages of reprogramming and, therefore, controlling it more precisely may lead to a higher rate of iPSC production.

When studying early stages of reprogramming in more detail, single-cell transcriptomic analysis has shown that changes in the expression of chromatin remodellers are one of the key molecular processes that distinguish cells that will undergo successful reprogramming from those that will fail[116] (FIG. 4B) This is consistent with previous population studies, where the importance of chromatin remodeling during the first stages of reprogramming had been highlighted (reviewed in [121]). Once cells become partially reprogrammed during the intermediate stages of the process, non-coding RNAs appear to regulate some of the reprogramming processes, namely by inducing metabolic changes as well as suppressing lineage- specific genes, thus suppressing the cell's differentiated phenotype[111] (FIG. 4B). At this stage, pluripotency genes, such as endogenous *Oct4*, also start to be expressed; however, their expression does not seem to be predictive of successful reprogramming until later on[113,115], indicating that the relatively early expression of these factors is not sufficient for iPSC reprogramming. Once the cells enter the late stages of reprogramming, pluripotency factors gain importance. For instance, sc-qPCR has suggested that *Sox2* drives a hierarchical sequence of gene expression changes[115], which finally leads to the cells reaching the iPSC state (FIG. 4B). This sequence of events has not only been studied at the transcriptome level; single-cell high resolution imaging has shown that more local transcriptional changes such as X chromosome reactivation, also occur in a stepwise manner[122].

Future perspective and conclusion

One of the greatest challenges of single-cell transcriptomics is the absence of spatial and temporal resolution, which is critically relevant in the context of developmental biology. Computational inferences, such as pseudotime ordering, have contributed to our understanding of defined differentiation trajectories by highlighting potential novel relevant factors, and have allowed the analysis of an extensive number of genes compared to current live imaging technologies. However, they are not sufficient to unravel the intricate mechanisms underlying early mammalian development. This is especially relevant once the epiblast and the primitive endoderm are established, as the embryo rapidly acquires a high level of complexity that is orchestrated by dynamic morphogenetic movements, cell proliferation and changes in gene expression, which are difficult to comprehensively capture.

In the past few years, there has been significant progress in determining spatial positions from single-cell transcriptional data. The establishment of imaging strategies to obtain thousands of single-molecule measurements in single cells has provided us with both real spatial coordinates and quantitative data. Moreover, computational biology has developed algorithms to infer cell positions by reconstructing spatial reference maps from images obtained with fluorescence microscopy[123,124]. New sample collection techniques, such as tomo-seq or laser capture microdissection have also enabled the transcriptomic analysis of particular cells extracted from specific spatial positions[125,126]. Of note, the latter technique has been recently applied to mouse embryonic development and led to the production of a high resolution map of E7.0 mouse epiblast cells, which has improved our understanding of the different expression patterns occurring at this developmental stage[127]. Although these techniques are generating promising outcomes, they are laborious and thus require further optimization.

Another area where single-cell technologies are advancing is in the development of combined methods to obtain as much information as possible from a single cell. This has currently been implemented at the level of chromatin accessibility, DNA methylation status and transcriptomic profile[128,129] and is expected to give us

a mechanistic view of the regulatory systems that govern developmental processes such as cell differentiation or cell-state transitions. Due to the relatively recent establishment of these techniques, we expect that computational strategies to simultaneously analyze the data will improve in the near future. Moreover, by eliminating the need for reverse transcription and non-specific amplification, direct sequencing of RNA, which is currently under development for the new generation of nanopore sequencers also has substantial potential to enhance future scRNA-seq protocols[130].

One of the main goals of single-cell transcriptomics remains the reconstruction of lineage trajectories using snapshots of continuous developmental processes. Incorporation of lineage tracing strategies into scRNA-seq protocols would therefore provide invaluable evidence to support the lineage hierarchies postulated from scRNA-seq data analyses. Recent studies have attempted to address this challenge by using the CRISPR–Cas9 genome editing system to introduce into eukaryotic cells a barcode coupled to a 'scratchpad' sequence harboring multiple copies of the target sequence of a specific guide RNA[40,131]. Through successive cell divisions, the Cas9 nuclease will introduce stochastic mutations onto the scratchpad sequence, thereby 'recording' the cell's lineage history and enabling the reconstitution of lineage hierarchies in a given tissue. This powerful technology has been currently applied *in vivo* in zebrafish[131] and *in vitro* in mESC cultures[40]. Coupling this tool to scRNA-seq will take the application of single-cell transcriptomics in developmental biology to a whole new level.

Single-cell technologies have already given us many new insights into early mammalian development. Molecular heterogeneity in early embryos as well as in mESCs underscore their highly dynamic nature, but we still do not fully understand the basis for this heterogeneity. We envisage that, in addition to optimization of current methods and the establishment of new techniques, as protocols mature and become more user-friendly, their adoption by the wider developmental biology research community will accelerate and increase their impact.

References

1. Spitzer, M. H. & Nolan, G. P. Mass Cytometry: Single Cells, Many Features. *Cell* **165,** 780–791 (2016).

2. Herzenberg, L. A. *et al.* The History and Future of the Fluorescence Activated Cell Sorter and Flow Cytometry: A View from Stanford. *Clin. Chem.* **48,** 1819–1827 (2002).

3. Brady, G., Barbara, M. & Iscove, N. Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. *Methods Mol Cell Biol.* **2,** 17–23 (1990).

4. Higuchi, R., Dollinger, G., Walsh, P. S. & Griffith, R. Simultaneous Amplification and Detection of Specific DNA Sequences. *Nat. Biotechnol.* **10,** 413–417 (1992).

5. Higuchi, R., Fockler, C., Dollinger, G. & Watson, R. Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions. *Nat. Biotechnol.* **11,** 1026–1030 (1993).

6. Guo, G. *et al.* Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. *Dev. Cell* **18,** 675–685 (2010).

7. Hu, M. *et al.* Multilineage gene expression precedes commitment in the hemopoietic system. *Genes Dev.* **11,** 774–785 (1997).

8. Miyamoto, T. *et al.* Myeloid or Lymphoid Promiscuity as a Critical Step in Hematopoietic Lineage Commitment. *Dev. Cell* **3,** 137–147 (2002).

9. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13,** 845–848 (2016).

10. Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* **33,** 269–276 (2015).

11. Kurimoto, K. *et al.* An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res.* **34,** e42–e42 (2006).

12. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6,** 377–382 (2009).

13. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear

Amplification. *Cell Rep.* **2,** 666–673 (2012).

14. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17,** 77 (2016).

15. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10,** 1096–1098 (2013). **Report presenting the Smart-seq2 protocol for single-cell RNA-seq.**

16. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30,** 777–782 (2012).

17. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161,** 1187–1201 (2015).

18. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015). **This article presents Drop-seq, which is a droplet-based protocol to perform single-cell RNA-seq.**

19. Ibarra-Soria, X. *et al.* Defining murine organogenesis at single cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nat. Cell Biol.* https://doi.org/10.1038/s41556-017-0013-z (2018).

20. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11,** 163–166 (2014).

21. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10,** 1093–1095 (2013).

22. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

23. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21,** 1543–1551

(2011).

24. Kolodziejczyk, A. A. *et al.* Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell* **17,** 471–485 (2015).

25. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14,** 309–315 (2017).

26. Maaten, L. J. P. van der & Hinton, G. E. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605 (2008).

27. Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 7426–7431 (2005).

28. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31,** 2989–2998 (2015).

29. Scialdone, A. *et al.* Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535**, 289-293 (2016). **This study characterizes the E6.5 epiblast, the Flk1^{+} mesodermal progenitor population and the differentiation path towards blood during gastrulation using the Smart-Seq2 protocol.**

30. Brunskill, E. W. *et al.* A gene expression atlas of early craniofacial development. *Dev. Biol.* **391,** 133–146 (2014).

31. DeLaughter, D. M. *et al.* Single-Cell Resolution of Temporal Gene Expression during Heart Development. *Dev. Cell* **39,** 480–490 (2016).

32. Petropoulos, S. *et al.* Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* **165,** 1012–1026 (2016).

33. Setty, M. *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637-45 (2016).

34. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–386 (2014).

35. Bendall, S. C. *et al.* Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell* **157,** 714–725 (2014).

36. Lubeck, E. & Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods* **9,** 743–748 (2012).

37. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11,** 360–361 (2014).

38. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA

molecules using multiple singly labeled probes. *Nat. Methods* **5,** 877–879 (2008).

39. Singer, Z. S. *et al.* Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells. *Mol. Cell* **55,** 319–331 (2014). **This study describes the expression dynamics of Nanog in mESCs, which includes transcription bursts.**

40. Frieda, K. L. *et al.* Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541,** 107–111 (2017).

41. Lee, J. H. *et al.* Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10,** 442–458 (2015).

42. Haim, L., Zipor, G., Aronov, S. & Gerst, J. E. A genomic integration method to visualize localization of endogenous mRNAs in living yeast. *Nat. Methods* **4,** 409–412 (2007).

43. Hocine, S., Raymond, P., Zenklusen, D., Chao, J. A. & Singer, R. H. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. *Nat. Methods* **10,** 119–121 (2013).

44. Graf, T. & Enver, T. Forcing cells to change lineages. *Nature* **462,** 587–594 (2009).

45. Hoppe, P. S. *et al.* Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature* **535,** 299–302 (2016).

46. Strumpf, D. *et al.* Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. *Development* **132,** 2093–2102 (2005).

47. Palmieri, S. L., Peter, W., Hess, H. & Schöler, H. R. Oct-4 Transcription Factor Is Differentially Expressed in the Mouse Embryo during Establishment of the First Two Extraembryonic Cell Lineages Involved in Implantation. *Dev. Biol.* **166,** 259–267 (1994).

48. Nichols, J. *et al.* Formation of Pluripotent Stem Cells in the Mammalian Embryo Depends on the POU Transcription Factor Oct4. *Cell* **95,** 379–391 (1998).

49. Chambers, I. *et al.* Functional Expression Cloning of Nanog, a Pluripotency Sustaining Factor in Embryonic Stem Cells. *Cell* **113,** 643–655 (2003).

50. Mitsui, K. *et al.* The Homeoprotein Nanog Is Required for Maintenance of Pluripotency in Mouse Epiblast and ES Cells. *Cell* **113,** 631–642 (2003).

51. Schrode, N., Saiz, N., Di Talia, S. & Hadjantonakis, A.-K. GATA6 Levels Modulate Primitive Endoderm Cell Fate Choice and Timing in the Mouse Blastocyst. *Dev. Cell* **29,** 454–467 (2014).

52. Dietrich, J.-E. & Hiiragi, T. Stochastic patterning in the mouse pre-implantation embryo. *Development* **134,** 4219–4231 (2007).

53. Goolam, M. *et al.* Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell* **165,** 61–74 (2016).

54. Alarcón, V. B. & Marikawa, Y. Deviation of the Blastocyst Axis from the First Cleavage Plane Does Not Affect the Quality of Mouse Postimplantation Development. *Biol. Reprod.* **69,** 1208–1212 (2003).

55. Hiiragi, T. & Solter, D. First cleavage plane of the mouse egg is not predetermined but defined by the topology of the two apposing pronuclei. *Nature* **430,** 360–364 (2004).

56. Motosugi, N., Bauer, T., Polanski, Z., Solter, D. & Hiiragi, T. Polarity of the mouse embryo is established at blastocyst and is not prepatterned. *Genes Dev.* **19,** 1081–1092 (2005).

57. Anani, S., Bhat, S., Honma-Yamanaka, N., Krawchuk, D. & Yamanaka, Y. Initiation of Hippo signaling is linked to polarity rather than to cell position in the pre-implantation mouse embryo. *Development* **141,** 2813–2824 (2014).

58. Bischoff, M., Parfitt, D.-E. & Zernicka-Goetz, M. Formation of the embryonic-abembryonic axis of the mouse blastocyst: relationships between orientation of early cleavage divisions and pattern of symmetric/asymmetric divisions. *Development* **135,** 953–962 (2008).

59. Johnson, M. H. & Ziomek, C. A. The foundation of two distinct cell lineages within the mouse morula. *Cell* **24,** 71–80 (1981).

60. Korotkevich, E. *et al.* The Apical Domain Is Required and Sufficient for the First Lineage Segregation in the Mouse Embryo. *Dev. Cell* **40,** 235–247.e7 (2017).

61. Piotrowska, K. & Zernicka-Goetz, M. Role for sperm in spatial patterning of the early mouse embryo. *Nature* **409,** 517–521 (2001).

62. Plachta, N., Bollenbach, T., Pease, S., Fraser, S. E. & Pantazis, P. Oct4 kinetics predict cell lineage patterning in the early mammalian embryo. *Nat. Cell Biol.* **13,** 117–123 (2011).

63. Tabansky, I. *et al.* Developmental Bias in Cleavage-Stage Mouse Blastomeres. *Curr. Biol.* **23,** 21–31 (2013).

64. Torres-Padilla, M.-E., Parfitt, D.-E., Kouzarides, T. & Zernicka-Goetz, M. Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature* **445,** 214–218 (2007).

65. Zernicka-Goetz, M. Development: Do Mouse Embryos Play Dice? *Curr. Biol.* **23,** R15–R17 (2013).

66. Biase, F., Cao, X. & Zhong, S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res.* **24**, 1787-96 (2014).

67. Shi, J. *et al.* Dynamic transcriptional symmetry-breaking in pre-implantation mammalian embryo development revealed by single-cell RNA-seq. *Development* **142,** 3468–3477 (2015).

68. Flach, G., Johnson, M. H., Braude, P. R., Taylor, R. A. & Bolton, V. N. The transition from maternal to embryonic control in the 2-cell mouse embryo. *EMBO J.* **1,** 681–686 (1982).

69. Blakeley, P. *et al.* Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142,** 3151–3165 (2015).

70. Braude, P., Bolton, V. & Moore, S. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature* **332,** 459–461 (1988).

71. Niakan, K. K. & Eggan, K. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Dev. Biol.* **375,** 54–64 (2013).

72. Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500,** 593–597 (2013).

73. Maître, J.-L. *et al.* Asymmetric division of contractile domains couples cell positioning and fate specification. *Nature* **536,** 344 (2016).

74. Maître, J.-L. Mechanics of blastocyst morphogenesis. *Biol. Cell* **109,** 323–338 (2017).

75. Chan, C. J., Heisenberg, C.-P. & Hiiragi, T. Coordination of Morphogenesis and Cell-Fate Specification in Development. *Curr. Biol.* **27,** R1024–R1035 (2017).

76. Chazaud, C., Yamanaka, Y., Pawson, T. & Rossant, J. Early Lineage Segregation between Epiblast and Primitive Endoderm in Mouse Blastocysts through the Grb2-MAPK Pathway. *Dev. Cell* **10,** 615–624 (2006).

77. Morris, S. A. *et al.* Origin and formation of the first two distinct cell types of the inner cell mass in the mouse embryo. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 6364–6369 (2010).

78. Morris, S. A., Graham, S. J. L., Jedrusik, A. & Zernicka-Goetz, M. The differential response to Fgf signalling in cells internalized at different times influences lineage segregation in preimplantation mouse embryos. *Open Biol.* **3,** 130104 (2013).

79. Yamanaka, Y., Lanner, F. & Rossant, J. FGF signal-dependent segregation of primitive endoderm and epiblast in the mouse blastocyst. *Development* **137,** 715–724 (2010).

80. Ohnishi, Y. *et al.* Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nat. Cell Biol.* **16,** 27–37 (2014). **Using single-cell transcriptomics, this study identifies *Fgf4* as one of the first genes to be differentially expressed within the ICM. Evaluation of *Fgf4* mutant embryos with scRNA-seq shows that *Fgf4*$^{-/-}$ cells are arrested before the decision between epiblast and primitive endoderm occurs.**

81. Xenopoulos, P., Kang, M., Puliafito, A., Di Talia, S. & Hadjantonakis, A.-K. Heterogeneities in Nanog Expression Drive Stable Commitment to Pluripotency in the Mouse Blastocyst. *Cell Rep.* **10,** 1508–1520 (2015).

82. Frankenberg, S. *et al.* Primitive Endoderm Differentiates via a Three-Step Mechanism Involving Nanog and RTK Signaling. *Dev. Cell* **21,** 1005–1013 (2011).

83. Molotkov, A., Mazot, P., Brewer, J. R., Cinalli, R. M. & Soriano, P. Distinct Requirements for FGFR1 and FGFR2 in Primitive Endoderm Development and Exit from Pluripotency. *Dev. Cell* **41,** 511–526.e4 (2017).

84. Kang, M., Garg, V. & Hadjantonakis, A.-K. Lineage Establishment and Progression within the Inner Cell Mass of the Mouse Blastocyst Requires FGFR1 and FGFR2. *Dev. Cell* **41,** 496–510.e5 (2017).

85. Arnold, S. J. & Robertson, E. J. Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nat. Rev. Mol. Cell Biol.* **10,** 91–103 (2009).

86. Sutherland, A. E. Tissue morphodynamics shaping the early mouse embryo. *Semin. Cell Dev. Biol.* **55,** 89–98 (2016).

87. Tam, P. P. L. & Behringer, R. R. Mouse gastrulation: the formation of a mammalian body plan. *Mech. Dev.* **68,** 3–25 (1997).

88. Tam, P. P., Parameswaran, M., Kinder, S. J. & Weinberger, R. P. The allocation of epiblast cells to the embryonic heart and other mesodermal lineages: the role of ingression and tissue movement during gastrulation. *Development* **124,** 1631–1642 (1997).

89. Wen, J. *et al.* Single-cell analysis reveals lineage segregation in early post-implantation mouse embryos. *J. Biol. Chem.* **292**, 9840-9854 (2017). **Using single cell transcriptional profiling, this study reveals the existence of a population of mesendodermal cells as early as E5.5, potentially one of the**

earliest populations after the exit from epiblast.

90. Rodaway, A. & Patient, R. Mesendoderm: An Ancient Germ Layer? *Cell* **105,** 169–172 (2001).

91. Tada, S. *et al.* Characterization of mesendoderm: a diverging point of the definitive endoderm and mesoderm in embryonic stem cell differentiation culture. *Development* **132,** 4363–4374 (2005).

92. Ema, M. *et al.* Combinatorial effects of Flk1 and Tal1 on vascular and hematopoietic development in the mouse. *Genes Dev.* **17,** 380–393 (2003).

93. Motoike, T., Markham, D. W., Rossant, J. & Sato, T. N. Evidence for novel fate of Flk1+ progenitor: contribution to muscle lineage. *Genes* **35,** 153–159 (2003).

94. Yamashita, J. *et al.* Flk1-positive cells derived from embryonic stem cells serve as vascular progenitors. *Nature* **408,** 92–96 (2000).

95. Ferdous, A. *et al.* Nkx2–5 transactivates the Ets-related protein 71 gene and specifies an endothelial/endocardial fate in the developing embryo. *Proc. Natl. Acad. Sci.* **106,** 814–819 (2009).

96. Rasmussen, T. L. *et al.* ER71 directs mesodermal fate decisions during embryogenesis. *Development* **138,** 4801–4812 (2011).

97. Gong, W. *et al.* Dpath software reveals hierarchical haemato-endothelial lineages of Etv2 progenitors based on single-cell transcriptome analysis. *Nat. Commun.* **8,** 14362 (2017).

98. Saga, Y. *et al.* MesP1 is expressed in the heart precursor cells and required for the formation of a

single heart tube. *Development* **126,** 3437–3447 (1999).

99. Devine, W. P., Wythe, J. D., George, M., Koshiba-Takeuchi, K. & Bruneau, B. G. Early patterning and specification of cardiac progenitors in gastrulating mesoderm. *eLife* **3,** e03848 (2014).

100. Lescroart, F. *et al.* Early lineage restriction in temporally distinct populations of Mesp1 progenitors during mammalian heart development. *Nat. Cell Biol.* **16,** 829–840 (2014).

101. Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292,** 154–156 (1981).

102. Weinberger, L., Ayyash, M., Novershtern, N. & Hanna, J. H. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nat. Rev. Mol. Cell Biol.* **17,** 155–169 (2016).

103. Guo, G. *et al.* Serum-Based Culture Conditions Provoke Gene Expression Variability in Mouse Embryonic Stem Cells as Revealed by Single-Cell Analysis. *Cell Rep.* **14,** 956–965 (2016).

104. Martinez Arias, A. & Brickman, J. M. Gene expression heterogeneities in embryonic stem cell populations: origin and function. *Curr. Opin. Cell Biol.* **23,** 650–656 (2011).

105. Filipczyk, A. *et al.* Network plasticity of pluripotency transcription factors in embryonic stem cells. *Nat. Cell Biol.* **17,** 1235–1246 (2015).

106. Cannon, D., Corrigan, A. M., Miermont, A., McDonel, P. & Chubb, J. R. Multiple cell and population-level interactions with mouse embryonic stem cell heterogeneity. *Development* **142,** 2840–2849 (2015).

107. Kalmar, T. *et al.* Regulated Fluctuations in Nanog Expression Mediate Cell Fate Decisions in Embryonic Stem Cells. *PLOS Biol.* **7,** e1000149 (2009).

108. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487,** 57–63 (2012). **Using single-cell imaging, this study identifies 2C-like cells, a rare subpopulation within mESCs that resembles the *in vivo* 2-cell stage.**

109. Zalzman, M. *et al.* Zscan4 regulates telomere elongation and genomic stability in ES cells. *Nature* **464,** 858–863 (2010).

110. Eckersley-Maslin, M. A. *et al.* MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. *Cell Rep.* **17,** 179–192 (2016). **This study shows similarities between the *in vivo* 2-cell state and 2C-like cells at the transcriptional level. Furthermore, 2C-like cells have open chromatin and hypomethylated DNA, both characteristics of the *in vivo* 2-cell stage.**

111. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126,** 663–676 (2006).

112. Takahashi, K. & Yamanaka, S. A decade of transcription factor-mediated reprogramming to pluripotency.

Nat. Rev. Mol. Cell Biol. **17,** 183–193 (2016).

113. Lujan, E. *et al.* Early reprogramming regulators identified by prospective isolation and mass cytometry. *Nature* **521,** 352–356 (2015).

114. O'Malley, J. *et al.* High-resolution analysis with novel cell-surface markers identifies routes to iPS cells. *Nature* **499,** 88–91 (2013).

115. Buganim, Y. *et al.* Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell* **150,** 1209–1222 (2012).

116. Chung, K.-M. *et al.* Single Cell Analysis Reveals the Stochastic Phase of Reprogramming to Pluripotency Is an Ordered Probabilistic Process. *PLOS ONE* **9,** e95304 (2014).

117. Zunder, E. R., Lujan, E., Goltsev, Y., Wernig, M. & Nolan, G. P. A Continuous Molecular Roadmap to iPSC Reprogramming through Progression Analysis of Single-Cell Mass Cytometry. *Cell Stem Cell* **16,** 323–337 (2015).

118. Kim, D. H. *et al.* Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming. *Cell Stem Cell* **16,** 88–101 (2015).

119. Polo, J. M. *et al.* A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells. *Cell* **151,** 1617–1632 (2012). **Transcriptional analyses define the reprogramming towards iPSC as a two-step process, where DNA methylation changes occur late in reprogramming.**

120. Smith, Z. D., Nachman, I., Regev, A. & Meissner, A. Dynamic single-cell imaging of direct reprogramming reveals an early specifying event. *Nat. Biotechnol.* **28,** 521–526 (2010).

121. Apostolou, E. & Hochedlinger, K. Chromatin Dynamics during Cellular Reprogramming. *Nature* **502,** 462–471 (2013).

122. Pasque, V. *et al.* X Chromosome Reactivation Dynamics Reveal Stages of Reprogramming to Pluripotency. *Cell* **159,** 1681–1697 (2014).

123. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33,** 503–509 (2015).

124. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33,** 495–502 (2015).

125. Espina, V. *et al.* Laser-capture microdissection. *Nat. Protoc.* **1,** 586–603 (2006).

126. Junker, J. P. *et al.* Genome-wide RNA Tomography in the Zebrafish Embryo. *Cell* **159,** 662–675 (2014).

127. Peng, G. *et al.* Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. *Dev. Cell* **36,** 681–697 (2016). **This study presents a spatial transcriptome map of the E7.0 epiblast. This was achieved using laser capture microdissection and profiling pools of 20 cells by scRNA-seq while keeping track of their original locations.**

128. Clark, S. J. *et al.* Joint Profiling Of Chromatin Accessibility, DNA Methylation And Transcription In Single Cells. *bioRxiv* 138685 (2017). doi:10.1101/138685 **This study reports a combined method to obtain the transcriptome, chromatin accessibility and DNA methylation states of individual cells.**

129. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12,** 519–522 (2015).

130. Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8,** ncomms16027 (2017).

131. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353,** aaf7907 (2016). **This study presents a proof-of-principle experiment of lineage tracking at single-cell resolution. Using a genomic barcode harboring unique CRISPR–Cas9 target sites in a zebrafish fertilized egg gives rise to an accumulation of thousands of uniquely edited barcodes in the offspring cells of the adult fish, allowing to infer lineage relationships between all adult cells.**

Acknowledgements

B.P.S is funded by the Wellcome Trust 4-Year PhD programme in Stem Cell Biology and Medicine and the University of Cambridge. C.G. is funded by the Swedish Research Council. Research in the Göttgens laboratory is supported by programme grants from the Wellcome, CRUK and Bloodwise, and by a Wellcome Strategic Award to study cell fate decisions during gastrulation (105031/D/14/Z). The authors also gratefully acknowledge core support funding from the Wellcome to the Wellcome– Medical Research Council Cambridge Stem Cell Institute.

Author contributions

B.P.S. researched data for the article and wrote the article. C.G. contributed substantially to the discussion of the content. B.P.S., C.G. and B.G. reviewed and edited the manuscript before submission.

Boxes

Box 1. Mouse embryonic development.

Fertilization gives rise to a totipotent cell, the zygote, which will generate all the embryonic and supportive extra-embryonic tissues necessary for development. These processes start with the segregation of the embryonic and extra-embryonic lineages before the embryo implants into the uterus wall. Once the embryo is implanted, the embryonic part will further diversify into the different organs that will make the adult organism.

Pre-implantation

Following fertilization, the zygote undergoes a series of symmetric divisions until the 8-cell stage. The embryo then becomes compacted and this aggregation of cells is termed 'morula' (see the figure, part **a**). During the 8-to-16 and the 16-to-32 cell stage transitions, some cells will divide symmetrically and will become specified into the outer cell layer, called trophectoderm (TE), which will contribute to the placenta; other cells will undergo asymmetric divisions, where one of the daughters will become trophectoderm and the other will become part of the inner cell mass (ICM), which will give rise to the embryo proper. Following this first lineage specification, the ICM undergoes a second fate choice and becomes segregated into the epiblast (EPI) and the primitive endoderm (PE) at the 32-to-64 cell stage, the latter giving rise to the yolk sac. At the same time, the embryo also undergoes a process of cavitation, and the blastocoel will form. Finally, at the blastocyst stage, spatial cell rearrangement gives rise to epiblast cells surrounded by outer primitive endoderm

Post-implantation

Following implantation, which occurs at embryonic day 4.5 (E4.5), the embryo undergoes extensive proliferation coupled with morphogenetic cell movements. First, at E5.5, it adopts a cup-like shape (dashed rectangle; see the figure, part **b**), which is characteristic of mouse embryonic development. At this stage, the distal visceral endoderm (DVE) becomes specified. The DVE is thought to be a signaling center from which antagonists of the signaling factor Nodal, such as Cerberus (Cer) and Lefty1, are expressed and repress Nodal in the adjacent epiblast. Subsequently, through an ill-defined mechanism, the anterior visceral endoderm (AVE) is specified at the anterior side of the embryo, which also secretes Nodal antagonists that seem to contribute to the formation of the primitive streak at the posterior side. During gastrulation, epiblast cells at the primitive streak undergo an epithelial-to-mesenchymal transition and become mesoderm, which then migrates out and colonizes the different regions of the embryo, including the extra-embryonic yolk sac, and forms the allantois. The primitive streak extends distally as gastrulation progresses, and the cells that egress later and more anteriorly are thought to become endoderm and colonize other tissues, such as the gut. In parallel, anterior epiblast cells adopt an ectodermal fate and will eventually form part of tissues such as the brain or the skin. During this developmental progression, the embryo also undergoes other morphological

changes. Particularly, the amnion and the chorion, two membranes that delimit the two embryonic cavities, namely the amniotic and the exocoelomic cavities, respectively, are formed.

Box 2. Advantages and limitations of single-cell transcriptional methods.

Over the past few years, several single-cell methods have been developed to transcriptionally profile individual cells. Each method has advantages and limitations, and the choice of technique will depend on the biological question one would like to address.

Single-cell qPCR measures the expression levels of a defined set of genes in single cells.

Advantages

More sensitive than single-cell RNA sequencing (scRNA-seq), as it captures lowly-expressedgenes more reliably.

Limitations

Targeted analysis.

Allows the analysis of a limited number of genes per experiment.

Well-based scRNA-seq methods (for example, Smart-seq2 and CEL-seq2). A collection of single-cell transcriptomics techniques where individual cells are sorted into single wells and processed separately (see the figure).

Advantages

Genome-wide approach.

Low transcript 3'-end bias compared to droplet-based methods (see below).

Limitations

Requires isolating single cells.

Time consuming, as each processed plate contains only 96 cells.

Droplet-based scRNA-seq methods (for example, Drop-seq, inDrop). This group of recently-developed methods utilizes microfluidics to transcriptionally profile multiple cells genome-wide in a high-throughput manner. As the figure shows, single cells are combined with barcoded beads in oil emulsions (droplets) that are subsequently collected into one tube and processed together.

Advantages

Genome-wide approach.

High-throughput.

Limitations

High 3'-end bias.

Lower efficiency of capturing unique transcripts.

Higher risk of analyzing cell doublets.

Figure legends

Figure 1. Single-cell transcriptomics.

A. Bioinformatics pipeline used to explore single-cell transcriptomics data. Following the normalization of gene expression data, visualization approaches (left), which are based on dimensionality reduction techniques, and clustering (right) are applied on the data. Rows in the heatmap depict single genes; columns refer to single cells; colors represent gene expression levels, where orange is high and blue is low. Cells are clustered into groups A—C based on their gene expression profiles, and genes are grouped into groups X–Z based on their expression profiles

B. Pseudotime analysis. Pseudotime paths, which highlight the progression of the biological process, are inferred from the dataset (top), and the dynamics of differentially expressed genes is analyzed for each path (bottom). The top diagram is a visualization of the transcriptome data using diffusion maps and the paths that the pseudotime algorithm has inferred. The color gradient seen in the arrows (from red to blue or to yellow) highlight progression from the starting cells (red) to the end cells (blue or yellow). The bottom diagrams show the different trends of gene expression along each of the paths.

C. In single-molecule RNA fluorescent in situ hybridization (sm-FISH) labeling, a fluorophore-based barcode is assigned to a specific RNA sequence. In combinatorial labeling, each RNA transcript is targeted with multiple sequence-specific probes with different fluorophores (three in the example) that can be detected under the microscope using different channels. The identity of each molecule is extrapolated from the pre-established code. In sequential labeling, each RNA is targeted with multiple sequence-specific unicolor probes. Once the sample is imaged, the probes are digested with DNase I and the next batch of probes with different fluorophores (one per RNA) is hybridized. This will result in a sequential color-coding, where RNA identities will be assigned using the pre-established code. PCA, principal component analysis; t-SNE, t-distributed stochastic neighbour embedding.

Figure 2. Cell fate specification during mouse embryonic development.

A. The decision between the inner cell mass (ICM) and the trophectoderm (TE). Gene expression differences increase as development progresses. Colors depict gene expression profiles. In the case of the 8-cell, 16-cell and 32-cell stages, colors highlight the type of cell division: symmetric or asymmetric. Cells become molecularly heterogeneous at the 2-cell stage and the molecular segregation is amplified in the offspring, represented as a progression from light to dark colours along the stages. This molecular segregation leads the blue cells to undergo symmetric divisions and become trophectoderm while yellow cells divide asymmetrically and give rise to both epiblast and primitive endoderm.

B. The 'ordered' and 'random' models of the decision between epiblast (EPI) and primitive endoderm (PE). In the ordered model, the first cells to be internalized at the 16-cell stage will give rise to the epiblast and the later- internalized cells will become the primitive endoderm. In the random model, cells are not specified (purple) until the 32-cell stage.

C. The interaction between Nanog$^+$ and Gata6$^+$ cells through fibroblast growth factor (FGF) signaling. Nanog represses Gata6 and promotes *Fgf4* expression. The resulting Fgf4 emerging from Nanog$^+$ cells binds to both Fgfr2 and Fgfr1 receptors in a paracrine manner in Gata6$^+$ cells and induces the expression of the primitive endoderm factors Gata4 and Sox17. At the bottom of the diagram, the transcriptional status of these cells is described, with Nanog$^+$ cells being *Nanog*$^+$*Gata6*$^-$*Fgf4*$^+$*Fgfr2*$^-$*Fgfr1*$^+$ and Gata6$^+$ cells being *Nanog*$^-$*Gata6*$^+$*Fgf4*$^-$*Fgfr2*$^+$*Fgfr1*$^+$.

D. Molecular heterogeneity (left) precedes morphological changes (right) such as the emergence of the primitive streak. First, cells can be divided into *T*$^+$, *Foxa2*$^+$ and *T*$^+$*Foxa2*$^+$ populations. Next, the primitive streak, where these cells are, is formed.

Figure 3. Mesoderm differentiation during gastrulation.

A. Heterogeneity within the mesoderm. Flk1$^+$ progenitors are molecularly heterogeneous (represented by the different colors). If studied as a whole, these will give rise to blood (orange), endothelium (purple) and

smooth muscle (blue). However, this progenitor population could be composed of populations of unipotent progenitor cells.

B. First and second wave of blood differentiation. The box (left) highlights the region in the yolk sac of mouse embryos at E7.5 where the blood islands can be found. During the first wave of hematopoiesis (zoomed box; left), primitive erythrocytes are apparent in the blood islands at E7.5. During the second wave (zoomed box; right), at around E8.25, the hemogenic endothelium generates erythroid and myeloid progenitors (EMPs), which differentiate into definitive blood cells.

Figure 4. Pluripotent stem cells to understand cell states.

A. Hypothetic mechanisms regulating Nanog levels that could cause cell-fate switch. Expression levels after cell division refer to the daughter cells. Top left: No cell-fate switch. Here, Nanog levels constantly increase during the cell cycle until reaching level 2 in the y axis. After cell division, the levels go back to the initial level 1. Top right: Fate switch by increasing the transcription rate. A higher transcription rate will result in higher maximal Nanog levels compared to 'no cell-fate switch' cells (level 4 instead of 2). This results in daughter cells with higher Nanog levels and the effect is amplified through consecutive cell divisions, leading to cell-fate switch towards the naïve state when Nanog levels are high enough. Bottom: Fate switch by extending the cell cycle. An extension of the cell cycle allows Nanog levels to increase, which again results in gradual increase in Nanog levels in daughter cells during consecutive cell divisions.

B. Reprogramming of somatic cells to induced pluripotent stem cells (iPSCs). The diagram highlights the knowledge about reprogramming acquired using single-cell technologies. Upon induction of reprogramming with the OSKM factors (Oct4, SRY-box 2 (Sox2), Krüppel-like factor 4 (Klf4) and Myc), a stochastic molecular heterogeneity arises in the cell population. Some of the cells undergo fast proliferation, and these will progress to the next step of the reprogramming process. The next step consists of chromatin remodelling followed by repression of differentiated cell fates with non- coding RNAs (ncRNA). The second phase of reprogramming begins with the activation of endogenous *Sox2* expression, which triggers the activation of a hierarchical program of gene expression that finalizes the conversion of the somatic cell into an iPSC.

Glossary

Fluorescence-activated cell sorting. Flow cytometry method to analyze and sort single cells based on the expression of cell surface markers.

Microfluidics systems. Automated technologies based on the use of microminiaturized devices for mixing and manipulating low fluid volumes aiming to achieve multiplexing and high-throughput yields.

Blastocyst. Embryonic stage composed of inner ICM cells, a fluid-filled cavity called blastocoel, and outer trophectoderm cells.

Epiblast. Group of cells derived from the ICM that will give rise to the embryo proper.

Primitive endoderm. Group of cells derived from the ICM that contribute to extra-embryonic tissues such as the yolk sac.

Unique molecular identifiers. Short sequences that uniquely tag individual RNA molecules.

Highly variable genes. Genes with biologically highly-variable expression levels.

Dimensionality-reduction approaches. Methods used in high dimensional datasets, where each gene represents a dimension, to reduce the number of dimensions and elicit the visualization of the dataset's structure in a 2- or 3-dimensional plot.

Inner cell mass (ICM). Group of cells located inside the blastocyst that will give rise to the primitive endoderm and the epiblast.

Trophectoderm. Group of cells located on the outer part of the blastocyst that will become the supportive extra-embryonic tissues, such as the placenta.

Blastomeres. The cells resulting from the first divisions of the fertilized egg.

Bimodal distribution. In gene expression analyses, refers to a gene being either (a) highly expressed or

(b) not or lowly expressed, with small numbers of cells displaying intermediate levels. Cells can thus be divided into two subpopulations based on the expression levels of that particular gene.

Maternal-to-zygotic transition. Process occurring shortly after fertilization, where maternal RNA and proteins are degraded and the zygotic genome is activated and produces RNA and proteins.

Morula. Early embryonic stage where the embryo is composed of a symmetric ball of morphologically similar cells.

Salt-and-pepper. Group of ICM cells with heterogeneous expression of epiblast and primitive endoderm markers, where some cells express more epiblast markers while others express more primitive endoderm markers.

Gastrulation. Embryonic process, following implantation, where epiblast cells become specified into the three germ layers (ectoderm, mesoderm and endoderm).

Primitive streak. Morphological structure at the posterior side of the embryo, formed by the accumulation of cells. It is where epiblast cells will egress to become mesoderm or endoderm.

Mesendodermal progenitor. A cell that can give rise to either mesoderm or endoderm.

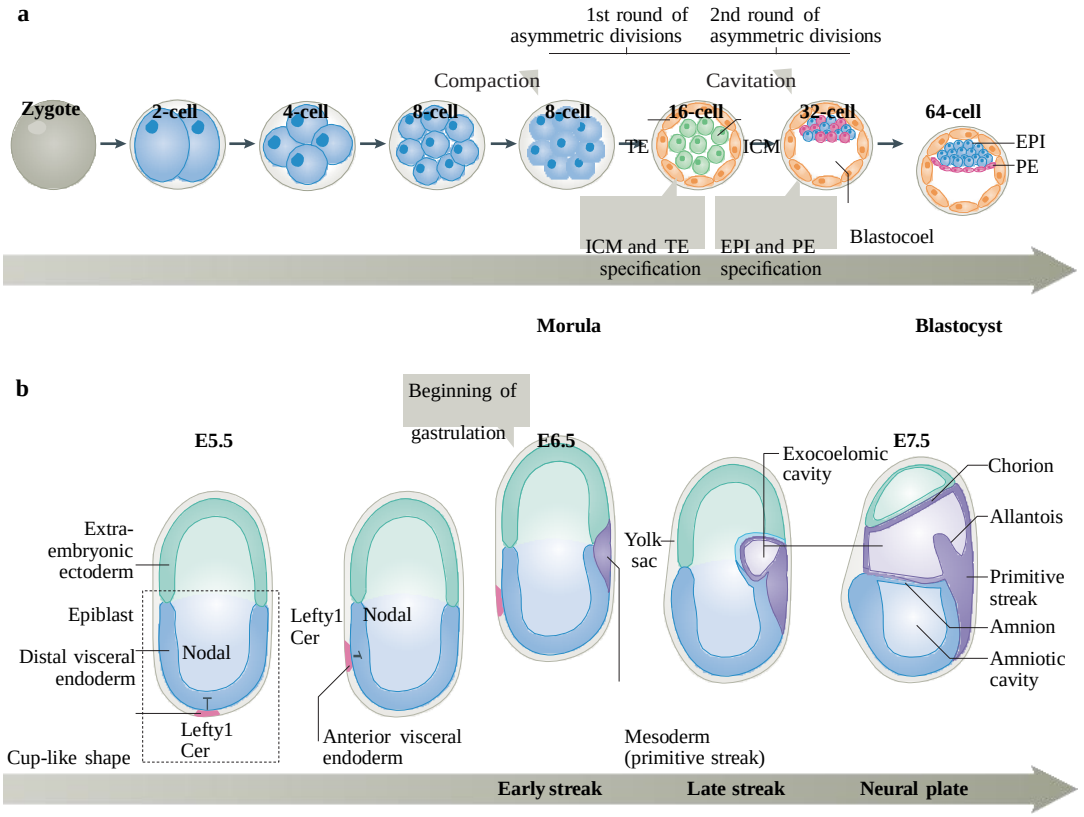
Yolk sac. Extra-embryonic tissue that originates from the primitive endoderm.

Boolean algorithm. Qualitative algorithm based on the Boolean (binary) logic, where only two values are accepted. In gene regulatory networks, one value will be 'active' and the other one 'inactive'.

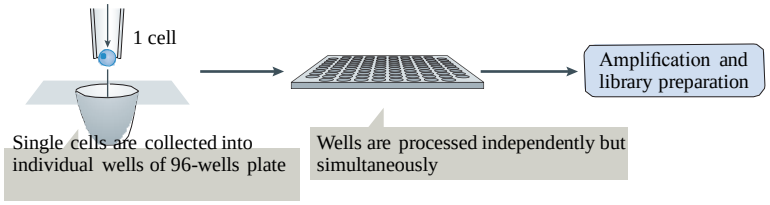
Unimodal distribution. In gene expression analyses, unimodal distribution refers to a gene being mostly expressed at intermediate levels.

X chromosome reactivation. Process where one of the X chromosomes in mammalian female cells becomes reactivated.

Box 1


a
1st round of asymmetric divisions
2nd round of asymmetric divisions
Compaction
Cavitation
Zygote
2-cell
4-cell
8-cell
8-cell
16-cell
32-cell
64-cell
TE
ICM
EPI
PE
Blastocoel
ICM and TE specification
EPI and PE specification
Morula
Blastocyst
b
Beginning of gastrulation
E5.5
E6.5
E7.5
Extra-embryonic ectoderm
Epiblast
Distal visceral endoderm
Nodal
Lefty1
Cer
Cup-like shape
Lefty1
Cer
Nodal
Anterior visceral endoderm
Yolk sac
Exocoelomic cavity
Mesoderm (primitive streak)
Chorion
Allantois
Primitive streak
Amnion
Amniotic cavity
Early streak
Late streak
Neural plate

Box 2 **Well-based**



Droplet-based

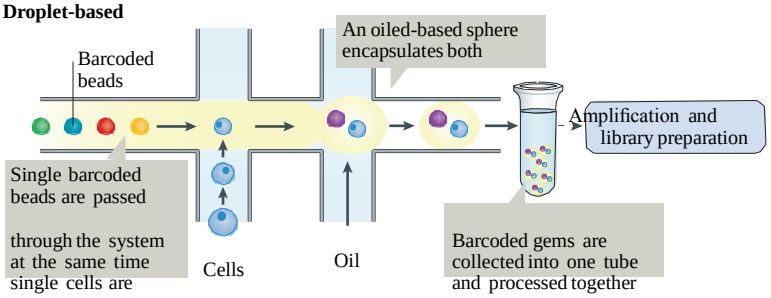


Fig 1

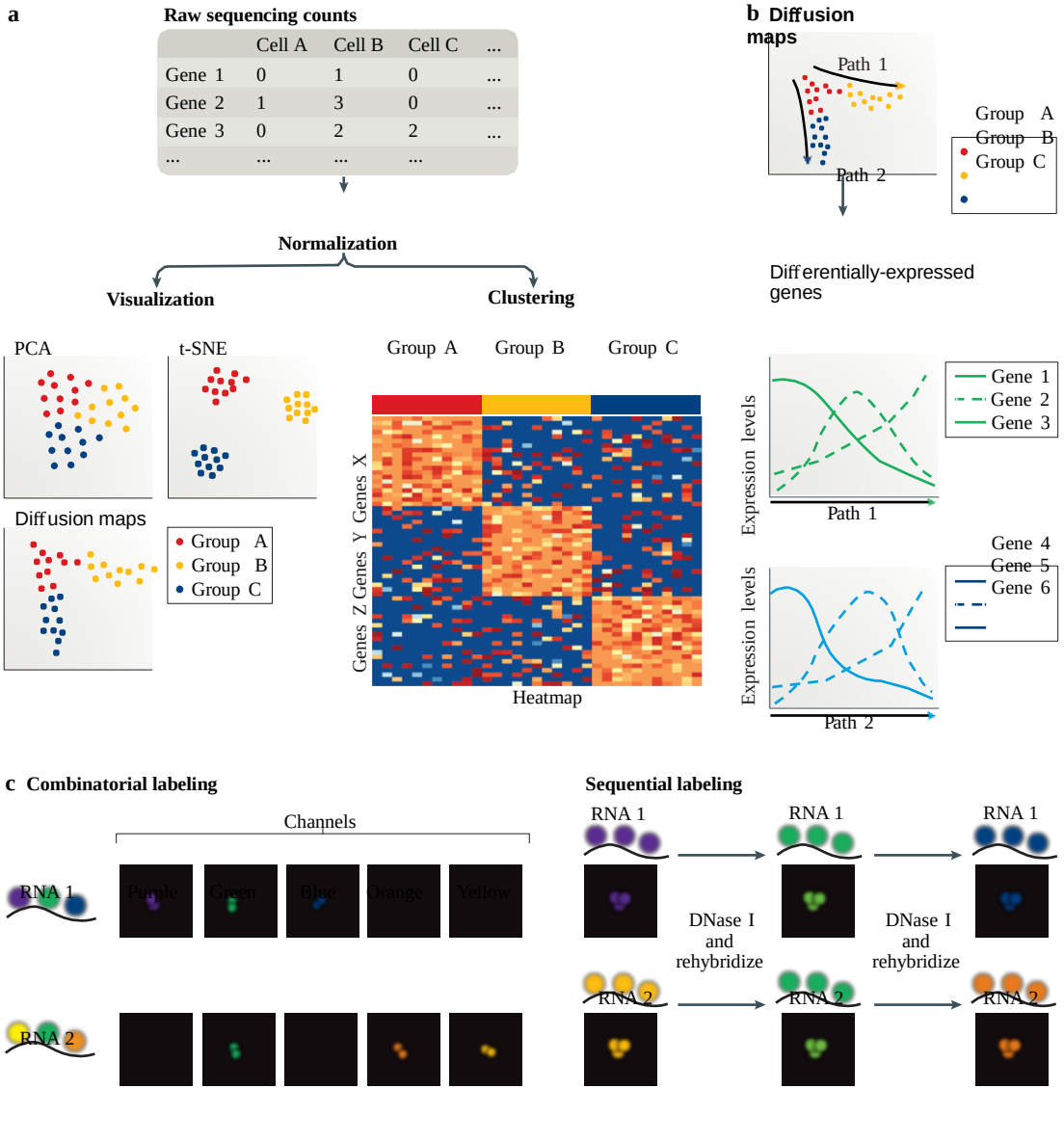


Fig 2

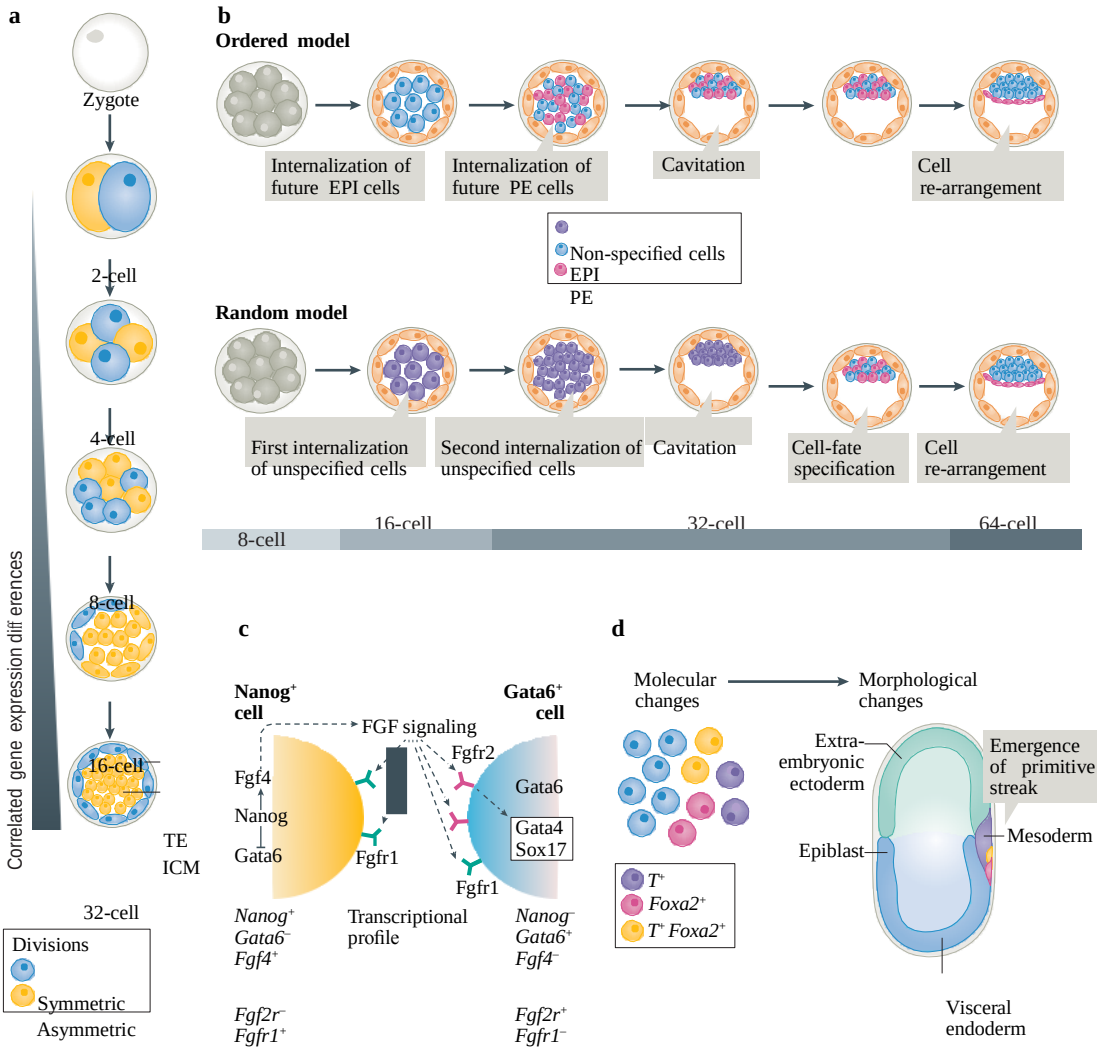

a
Zygote
2-cell
4-cell
8-cell
16-cell
32-cell
Correlated gene expression differences
TE
ICM
Divisions
Symmetric
Asymmetric
b
Ordered model
Internalization of future EPI cells
Internalization of future PE cells
Cavitation
Cell re-arrangement
Non-specified cells
EPI
PE
Random model
First internalization of unspecified cells
Second internalization of unspecified cells
Cavitation
Cell-fate specification
Cell re-arrangement
8-cell
16-cell
32-cell
64-cell
c
Nanog+ cell
Gata6+ cell
FGF signaling
Fgfr2
Fgf4
Nanog
Gata6
Fgfr1
Gata6
Gata4
Sox17
Fgfr1
Nanog+ Gata6− Fgf4+
Transcriptional profile
Nanog− Gata6+ Fgf4−
Fgf2r− Fgfr1+
Fgf2r+ Fgfr1−
d
Molecular changes
Morphological changes
T+
Foxa2+
T+ Foxa2+
Extra-embryonic ectoderm
Emergence of primitive streak
Epiblast
Mesoderm
Visceral endoderm

Fig 3

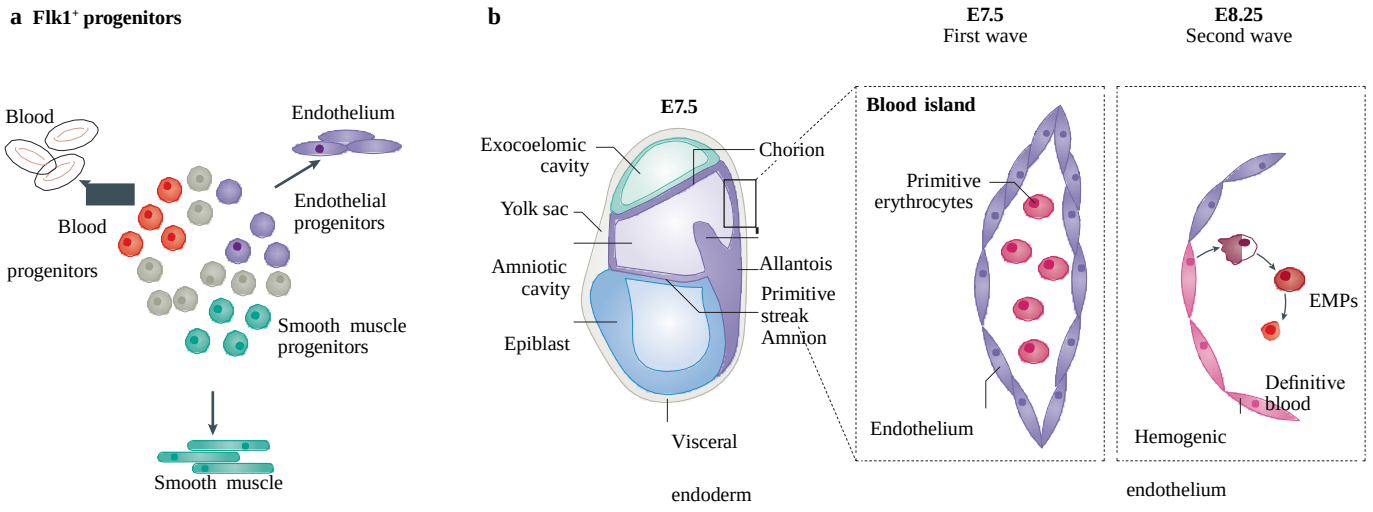
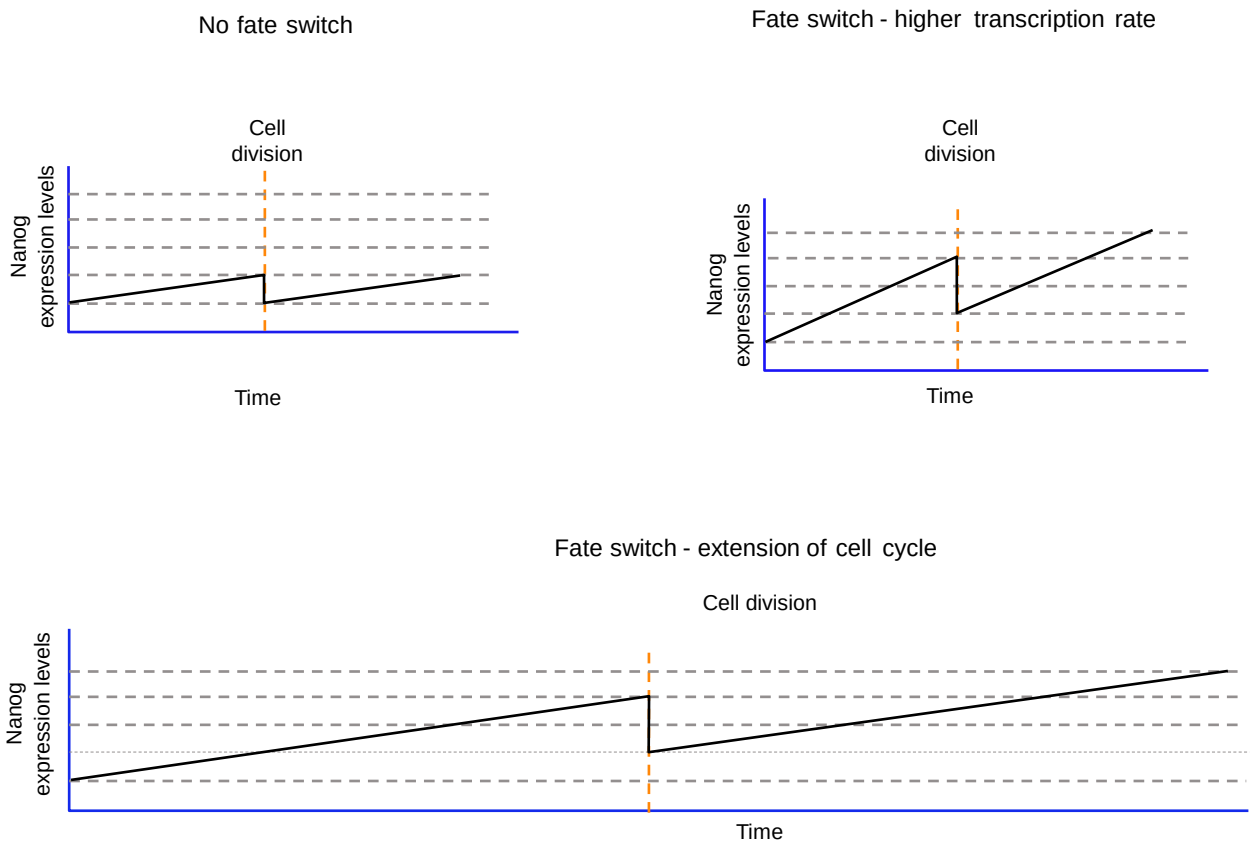
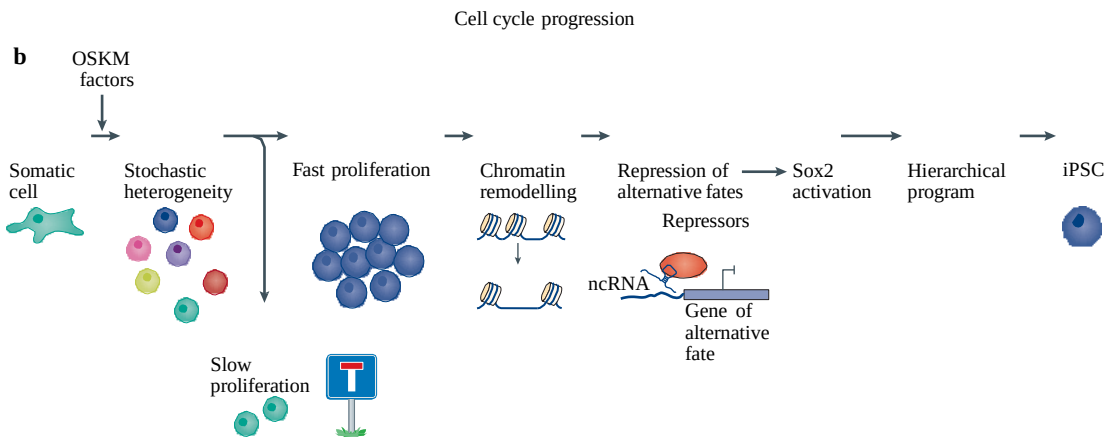

a Flk1+ progenitors
Blood
Endothelium
Endothelial progenitors
Blood progenitors
Smooth muscle progenitors
Smooth muscle
b
E7.5
First wave
E8.25
Second wave
E7.5
Exocoelomic cavity
Chorion
Yolk sac
Amniotic cavity
Allantois
Primitive streak
Epiblast
Amnion
Visceral endoderm
Blood island
Primitive erythrocytes
Endothelium
EMPs
Definitive blood
Hemogenic endothelium

Fig 4

a


No fate switch
Cell division
Nanog expression levels
Time
Fate switch - higher transcription rate
Cell division
Nanog expression levels
Time
Fate switch - extension of cell cycle
Cell division
Nanog expression levels
Time


Cell cycle progression
b
OSKM factors
Somatic cell
Stochastic heterogeneity
Fast proliferation
Slow proliferation
Chromatin remodelling
Repression of alternative fates
Repressors
ncRNA
Gene of alternative fate
Sox2 activation
Hierarchical program
iPSC